The best of both worlds: synthesis-based acceleration for physics-driven cosparse regularization

Srdan Kitic, Nancy Bertin, Rémi Gribonval

The best of both worlds: synthesis-based acceleration for physics-driven cosparse regularization

Srđan Kitić[1], Nancy Bertin[2] and Rémi Gribonval[3].

[1]Technicolor R&D, [2]CNRS - UMR 6074, [3]Inria, France. *

***Abstract*— Recently, a regularization framework for ill-posed inverse problems governed by linear partial differential equations has been proposed. Despite nominal equivalence between sparse synthesis and sparse analysis regularization in this context, it was argued that the latter is preferable from computational point of view (especially for huge scale optimization problems arising in physics-driven settings). However, the synthesis-based optimization benefits from simple, but effective all-zero initialization, which is not straightforwardly applicable in the analysis case. In this work we propose a multiscale strategy that aims at exploiting computational advantages of both regularization approaches.**

1 Introduction

Linear partial differential equations (pde) are ubiquitous in mathematical models of physical laws. Hence, whether in implicit or explicit form, they appear in various signal processing inverse problems (ranging from, *e.g.* sound source localization to brain imaging). Inspired by impressive results in sparse signal recovery and compressed sensing [1], several works *e.g.* [2, 3, 4, 5, 6] have proposed an explicit use of such physical models in regularization of some highly ill-posed inverse problems (baptized *"physics-driven"* regularization methods).

Generally, a linear pde models the relation among two physical quantities (x, z) as $Ax = z$, where the linear operator A encapsulates the pde with appropriate initial and/or boundary conditions. Analogously, one can write $x = Dz$, where D is a linear operator acting as an "inverse" to A. Particularly, D is the integral operator encoding the so-called Green's functions, *i.e.* impulse responses of the operator A. One is interested in inferring the quantity z, which is often characterized by a small number of free parameters (representing, for instance, dominant sources of brain activity in an EEG application). On the other hand, we are only given a few measurements y of the quantity x (*e.g.* voltage measurements at the surface of the head). The measurements are, therefore, acquired by applying a subsampling operator M to the signal x. This problem is severely ill-posed, and one way of addressing it is by asking for an estimate $\hat{z}$ (analogously, $A\hat{x}$) having the lowest complexity, *i.e.* the fewest degrees of freedom possible.

Analytical solutions of pdes are available only in certain restricted regimes. In other cases, one approaches the problem numerically and *discretizes* the involved quantities and operators ($A \to \mathbf{A} \in \mathbb{R}^{\mathsf{n}\times\mathsf{n}}$, $x \to \mathbf{x} \in \mathbb{R}^{\mathsf{n}}$, $D \to \mathbf{D} \in \mathbb{R}^{\mathsf{n}\times\mathsf{n}}$, $z \to \mathbf{z} \in \mathbb{R}^{\mathsf{n}}$, $M \to \mathbf{M} \in \mathbb{R}^{\mathsf{m}\times\mathsf{n}}$, $y \to \mathbf{y} \in \mathbb{R}^{\mathsf{m}}$). It should be clear that $\mathbf{D} = \mathbf{A}^{-1}$, which is identical to computing the response of a linear system defined by $\mathbf{A}$ for an impulse placed at every point of a discrete n-dimensional domain.

*This work was supported in part by the European Research Council, PLEASE project (ERC-StG-2011-277906).

Low complexity can be promoted through *sparsity* [1] of $\mathbf{z}$ (minimizing $\|\mathbf{z}\|_0$) or *cosparsity* [7] of $\mathbf{x}$ (minimizing $\|\mathbf{A}\mathbf{x}\|_0$). A common relaxation to these problems is the constrained ℓ_1 norm minimization (*a.k.a. basis pursuit*), either in the *sparse analysis*

$$\underset{\mathbf{x}}{\text{minimize}}\ \|\mathbf{A}\mathbf{x}\|_1 \ \text{subject to}\ \mathbf{M}\mathbf{x} = \mathbf{y}, \tag{1}$$

or *sparse synthesis* flavor

$$\underset{\mathbf{z}}{\text{minimize}}\ \|\mathbf{z}\|_1 \ \text{subject to}\ \mathbf{M}\mathbf{D}\mathbf{z} = \mathbf{y}. \tag{2}$$

The pde-encoding matrix $\mathbf{A}$ thus represents the *analysis operator*, while the row-reduced Green's function-encoding matrix $\mathbf{MD}$ represents the (synthesis) *dictionary*.

2 The Chambolle-Pock algorithm

A popular method for solving large scale nonsmooth problems such as (1) and (2) is the so-called *Chambolle-Pock* or *preconditioned ADMM* algorithm [8]. It is a primal-dual approach based on the *saddle point* interpretation of the original constrained problem. Iteratively solving intermediate primal and dual problems avoids matrix inversion, hence its per-iteration cost is dominated by the cost of evaluating matrix-vector products and proximal operators. To make the latter efficient, one needs to appropriately customize the saddle-point problem to leverage all available structures.

Particularly, in the analysis case, we exploit the fact that $\mathbf{M}$ is a row-reduced identity matrix. This allows for cheap projection to a set $\Theta = \{\mathbf{x} \mid \mathbf{M}\mathbf{x} = \mathbf{y}\}$, leading to the following saddle point formulation:

$$\underset{\mathbf{x}}{\text{minimize}}\ \underset{\boldsymbol{\lambda}}{\text{maximize}}\ \langle \mathbf{A}\mathbf{x}, \boldsymbol{\lambda}\rangle + \chi_\Theta(\mathbf{x}) - \ell_1^*(\boldsymbol{\lambda}), \tag{3}$$

where χ_Θ is the indicator function of the set Θ and ℓ_1^* is the convex conjugate [9, 11] of the ℓ_1 norm function (*i.e.* an indicator function of the ℓ_∞ ball). In the synthesis case, we exploit the separability of the $\|\mathbf{z}\|_1$ objective, which yields the standard Lagrangian problem:

$$\underset{\mathbf{z}}{\text{minimize}}\ \underset{\boldsymbol{\lambda}}{\text{maximize}}\ \langle \mathbf{M}\mathbf{D}\mathbf{z} - \mathbf{y}, \boldsymbol{\lambda}\rangle + \ell_1(\mathbf{z}). \tag{4}$$

In both cases, $\boldsymbol{\lambda}$ represents the corresponding dual variable.

The Chambolle-Pock algorithm essentially evaluates two proximal operators per iteration, each assigned to primal and dual variable, respectively. For the presented problems, the algorithm is actually (asymptotically) first-order optimal, since it obtains $\mathcal{O}(1/\mathsf{k})$ convergence rate[1] [8, 10] when all penalties are non-smooth, but structured [11]. More precisely, decrease of the primal-dual gap is proportional to $\|\mathbf{A}\|_2^2/\mathsf{k}$, in the analysis, and to $\|\mathbf{MD}\|_2^2/\mathsf{k}$, in the synthesis case ($\|\cdot\|_2$ denotes the induced 2-norm of a matrix).

[1]Where k denotes the iteration count.

3 Computational differences

Assuming that the regularization indeed yields well-posed problems, solving (1) or (2) is equivalent, in the sense that using the solution of one problem, we can easily recover the solution of another (since $\mathbf{Ax} = \mathbf{z}$). However, as demonstrated in [12], the two optimization problems significantly differ from computational point of view. In fact, if the applied discretization is locally supported (which is often the case with, *e.g.*, finite difference or finite element discretization methods), the analysis operator $\mathbf{A}$ is extremely sparse (with $\mathcal{O}(\mathsf{n})$ nonzero entries), while the dictionary $\mathbf{MD}$ is most often a dense matrix ($\mathcal{O}(\mathsf{mn})$). But the differences do not end there: as widely recognized [13], physical problems are often unstable, since small changes in z can induce large fluctuations of x. In line with that, discretization usually leads to increasingly ill-conditioned systems: the condition number $\kappa = \sigma_{\max}/\sigma_{\min}$ of $\mathbf{A}$ (*eq.* $\mathbf{D}$) grows fast with n. However, one can often factorize the analysis operator (with abuse of notation) as $\tau\mathbf{A}$, where the scale factor τ depends only on the discretization stepsize and the entries of $\mathbf{A}$ remain constant (this will be validated on the actual example in the following section). Notice that, in the basis pursuit problem (1), the scale τ does not affect the solution, and can be neglected. Now, the growth of κ is due to decrease of the smallest singular value of $\mathbf{A}$ (*i.e.* increase of the largest singular value of $\mathbf{D} = \mathbf{A}^{-1}$), hence $\sigma_{\max}(\mathbf{A}) = \|\mathbf{A}\|_2$ is stable.

The consequence for the primal-dual algorithm discussed previously, is that (at worst) the synthesis approach will require orders of magnitude more iterations to converge, in addition to high computational cost per iteration. Given these arguments, one may conclude that it should be completely avoided in the physics-driven context. However, it has an important advantage over the analysis-based optimization: since the expected solution is sparse, a simple all-zero initial estimate is already close to the optimal point. In order to exploit this feature, we propose a simple scheme: i) apply *crude* (low-resolution) discretization, and solve the problem (4) to obtain the crude estimate $\tilde{\mathbf{x}} = \mathbf{D}\tilde{\mathbf{z}}$; ii) interpolate $\tilde{\mathbf{x}}$ to a target high-resolution discretization and use it as an initial point $\hat{\mathbf{x}}^{(0)}$ for the problem (3).

4 An example: 1D Laplacian

We will demonstrate the idea on a simple one-dimensional problem. Assume that on a domain $\mathrm{r} \in [0, \phi]$, a physical process is modeled as

$$\frac{d^2x(\mathrm{r})}{d\mathrm{r}^2} = z(\mathrm{r}), \tag{5}$$

with $x(0) = x(\phi) = 0$ (*e.g.* modeling a potential distribution of a grounded thin rod, with sparse "charges" $z(\mathrm{r})$).

By applying the second order finite difference discretization to this problem, we end up with a well-known symmetric tridiagonal Toeplitz matrix $\mathbf{A}$ (*i.e.* 1D discrete Laplacian), with a "stencil" defined as $\tau[-1,\ 2,\ -1]$ (henceforth, we neglect τ). This simple matrix allows for fast computation of $\mathbf{A}^{-1}\mathbf{z}$ using the Thomas algorithm [14]. In addition, it admits simple analytical expressions for extremal singular values [15], namely $\sigma_{\max} = 2 - 2\cos(\frac{\mathsf{n}\pi}{\mathsf{n}+1})$, and $\sigma_{\min} = 2 - 2\cos(\frac{\pi}{\mathsf{n}+1})$. The ill-conditioning with regards to the size n is obvious, but the true value of $\|\mathbf{MD}\|_2$ is somewhat lower than $1/\sigma_{\min}$, since it also depends on the number of measurements m and the realization of the presumed random sampling. In general, one expects $\|\mathbf{MD}\|_2 \to 1/\sigma_{\min}$ as $\mathsf{m} \to \mathsf{n}$.

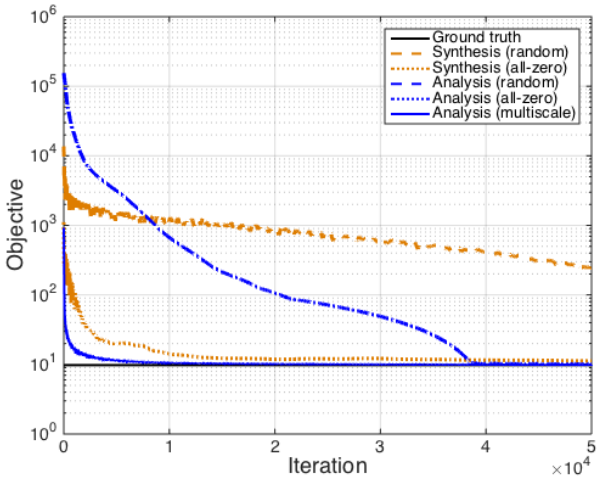

Figure 1: "Objective" $:= \|\mathbf{s}\|_1 + \|\mathbf{r}\|_2^2$, where $\mathbf{s}$ is a sparse estimate (*e.g.* $\mathbf{Ax}$) and $\mathbf{r}$ is a residual vector (*e.g.* $\mathbf{Mx} - \mathbf{y}$).

To verify our claims, we simulated the problem of size $\mathsf{n} = 10^3$, with the size of the support set being $\|\mathbf{z}\|_0 = 10$ (chosen uniformly at random from the set $[1, \mathsf{n}]$, with coefficients iid distributed as $\mathcal{N}(0, 1)$). The measurement vector $\mathbf{y}$ contains (selected uniformly at random) $\mathsf{m} = 250$ samples of the signal $\mathbf{x} = \mathbf{A}^{-1}\mathbf{z}$. The iteration threshold is set to $\mathsf{k}_{\max} = 5 \times 10^4$.

We first solve both problems (1) and (2) (using the appropriate versions of the Chambolle-Pock algorithm), by generating the initial points $\hat{\mathbf{x}}^{(0)}$ and $\hat{\mathbf{z}}^{(0)}$ randomly (iid sampled from $\mathcal{N}(0, 1)$), and then re-running both versions with all-zero initialization. Objective function decrease graphs in Figure 1 confirm our findings: when both algorithms are randomly initialized, the analysis one exhibits considerably faster convergence (moreover, the synthesis version does not reach the ground truth value). However, when the synthesis algorithm is initialized with an all-zero vector, it converges rapidly, outperforming the analysis approach in both cases (for which, interestingly, the two initializations yield the same convergence curve).

Unfortunately, in practice we are rarely capable of efficiently applying the synthesis approach, since cheap computation of $\mathbf{A}^{-1}\mathbf{z}$ is possible only in specific cases. Otherwise, $\mathbf{MD}$ needs to be explicitly computed and stored, leading to memory bottlenecks and high per-iteration cost. To alleviate this issue, we exploit the aforementioned multiscale strategy. First, the synthesis version is appropriately initialized (with an all-zero vector) and solved on a crude, $\mathsf{n}_{\mathsf{low}} = 500$ grid. Then its spline-interpolated (*cf.* [15]) estimate is used to initialize the full resolution analysis-based solver. The "analysis (multiscale)" graph presented in Figure 1 verifies that this scheme is indeed very efficient, in this case converging the fastest among all considered algorithms and initializations.

5 Conclusion

We have presented a simple, yet effective acceleration of the analysis-based optimization, in the physics-driven setting. Leveraging the synthesis-based initialization enables orders of magnitude faster convergence compared to the naive case. Even though only a simple 1D Laplacian case was discussed and justified, we feel that the same methodology holds in more involved scenarios, comprising different multidimensional pdes with complicated boundary conditions.

References

[1] S. Foucart and H. Rauhut, "A mathematical introduction to compressive sensing", Springer, **1**(3), 2013.

[2] D. Malioutov, M. Çetin and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays", IEEE Transactions on Signal Processing, **53**(8):3010–3022, 2005.

[3] L. Stenbacka, S. Vanni, K. Uutela and R. Hari, "Comparison of minimum current estimate and dipole modeling in the analysis of simulated activity in the human visual cortices", NeuroImage, **16**(4): 936–943, 2002.

[4] I. Dokmanić and M. Vetterli, "Room helps: Acoustic localization with finite elements", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 2617–2620, 2012.

[5] S. Nam and R. Gribonval, "Physics-driven structured cosparse modeling for source localization", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 5397–5400, 2012.

[6] J. Murray-Bruce and P. L. Dragotti, "Spatio-temporal sampling and reconstruction of diffusion fields induced by point sources", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 31–35, 2014.

[7] S. Nam, M. E. Davies, M. Elad and R. Gribonval, "The cosparse analysis model and algorithms", Applied and Computational Harmonic Analysis, **34**(1):30–56, 2013.

[8] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging", Journal of Mathematical Imaging and Vision, **40**(1): 120–145, 2011.

[9] S. Boyd and L. Vandenberghe, "Convex optimization", Cambridge university press, 2004.

[10] A. Chambolle and T. Pock, "On the ergodic convergence rates of a first-order primal–dual algorithm", Mathematical Programming (Springer): 1–35, 2015.

[11] Y. Nesterov, "Introductory lectures on convex optimization: A basic course", Springer Science & Business Media, **87**, 2013.

[12] S. Kitić, L. Albera, N. Bertin and R. Gribonval, "Physics-driven inverse problems made tractable with cosparse regularization", IEEE Transactions on Signal Processing, **64**(2): 335–348, 2016.

[13] V. Isakov, "Inverse problems for partial differential equations", Springer Science & Business Media, **127**, 2006.

[14] L. H. Thomas, "Using a computer to solve problems in physics", Applications of Digital Computers (Boston: Ginn and Company): 44–45, 1963.

[15] G. Strang, "Computational science and engineering", Wellesley-Cambridge Press Wellesley, **1**, 2007.